DOCUMENT RESUME

ED 167 569 TM 007 677

AUTHOR	Folsom, R. E.; Weber, J. H.
TITLE	Limited Investigation of Effects of Stratification, Clustering, Respondent Mobility, and Overlapping Schools with the 1972 Sample. 220-884-38.
INSTITUTION	Research Triangle Inst., Research Triangle Park, N.C. Center for Sampling Research and Design.
PUB DATE	Apr 76
NOTE	31p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Followup Studies; Graduate Surveys; High School Graduates; *National Surveys; Reliability; *Research Design; *Sampling; Senior High Schools; Statistical Analysis; *Statistical Bias; Student Mobility
IDENTIFIERS	*National Longitudinal Study High School Class 1972

ABSTRACT

Longitudinal measures of change are the most important kinds of estimates being produced from the National Longitudinal Study of the High School Class of 1972 (NLS). Several topics are presented in this report, most of which build on previous variance components analyses. The major areas covered are effects of stratification and clustering, correlations from overlapping schools in the two NLS surveys, and mobility of the NLS 1972 cohort. The effects of clustering students and stratifying schools, based on the change-variable variance components, are discussed. This methodology parallels that developed by Moore and Shah for base-year statistics. The effects of retaining some of the previous cohort sample schools for the new cohort and the advantages of a fifty percent overlap are also described. The mobility of the first cohort of sample students has direct cost implications for followup surveys. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED167569

ZZU-884-38

LIMITED INVESTIGATION OF EFFECTS OF STRATIFICATION, CLUSTERING, RESPONDENT MOBILITY, AND OVERLAPPING SCHOOLS WITH THE 1972 SAMPLE.

by

R. E. Folsom
J. H. Weber

Sampling Research and Design Center
Research Triangle Institute
Research Triangle Park, North Carolina 27709

April 1976

TM007 627

NLS LIBRARY

RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

Table of Contents

	Page
List of Tables	ii
I. Introduction	1
II. Effects of Stratification and Clustering	5
III. Correlations from Overlapping Schools in the Two NLS Surveys	12
IV. Mobility of the NLS 1972 Cohort	23
References	26

List of Tables

Page

Table 1: Average effects of clustering and stratification for the NLS design using base-year variance component estimates. 8

Table 2: Average effects of clustering and stratification for the NLS design using change-variable variance component estimates. 9

Table 3: Formulas for calculating the variance of the differences between two means. 15

Table 4: The percent reduction in the estimated variance when comparing means of two NLS cohorts due to overlapping schools using base-year variance component estimates. . . 17

Table 5: The percent reduction in the estimated variance when comparing means of two NLS cohorts due to overlapping schools using change-variable variance component estimates. 18

Table 6: Number of student records by matching category for BY, FFU, and SFU. 24

I. INTRODUCTION

Longitudinal measures of change are the most important kinds of estimates being produced from the NLS survey. Several topics are presented in this report, most of which build on the previous variance component analyses. These topics are introduced in the following paragraphs.

In a previous report by Moore and Shah [1], the effects of stratification, clustering, and unequal weighting on the variances of NLS statistics were discussed using the base-year variance component estimates. Because of the importance of longitudinal comparisons such as those contained in the base-year (BY) to first followup (FFU) change-variables, section II presents a design effect partitioning based on the change-variable variance components. The methodology developed in this section parallels that developed by Moore and Shah for base-year statistics. The effects of stratification and clustering on base-year statistics are compared here with the corresponding change-variable effects. Average design effects were computed over 14 change-variables for national means and eight subgroup means including whites, blacks, other races, females, males, and three levels of father's education. Effects for differences between subgroup and national means were also averaged. Finally, an over-all average across all types of statistics and subgroups was determined. These design effects were then partitioned into components due to stratification and clustering. The effect of unequal weighting is discussed briefly. The findings indicate that clustering had on the average increased the variance over all statistics by 23.6 percent for change-variable statistics compared with 39.1 percent for base-year statistics. Stratification, however, reduced the variance over all statistics by 9.8 percent

for the change-variable estimates in contrast to 14.3 percent for the change-variable statistics.

There is generally an advantage in selecting the same sampling units for two surveys from the same population at two time points when comparison between the two time points are important. Section III investigates the effects of retaining some of the NLS 1972 cohort sample schools for the new NLS cohort planned for 1978. The reduction in variance for first versus second cohort contrasts is computed assuming 50 percent and 25 percent of the sample schools are the same for both cohorts. The amount of variance reduction expected when contrasting school matched samples with two independent samples of schools depends on the magnitude of the positive correlation between the mean response of students within the same school for the two time periods. Three values of r, the correlation coefficient, were examined; namely r=1, .75, and .50. The r=1 value provides an upper bound for the amount of variance reduction while the lower values are more likely to approximate observed levels of correlation.

While the variance of between cohort contrasts would be minimized by 100 percent overlap, the precision of new NLS cohort estimates would suffer from school closings, grade range changes, and enrollment shifts. New high schools would also be missed, biasing the new cohort results. On the other hand, a 50 percent overlap scheme could be combined with a selection procedure developed by Keyfitz [3] which allows one to select a new sample using updated size measures (current senior enrollments) while maximizing the probability that a specified fraction of the old sample will be retained. Our projections indicate that with r no larger than .50 a modest six to nine percent reduction in the variance of national means

could be achieved with a 50 percent overlap.

The mobility of the 1972 cohort is discussed in section IV. Students are considered movers if their first or second followup addresses are located in a different county from their base-year address and the associated five digit zip codes are different. The mobility of sample students has direct costs implications for followup surveys. The more mobile the student population, the greater the cost associated with personal interview followup of mail nonrespondents. In previous reports by Moore and Shah [4] and Folsom and Weber [5], mobility rates ranging from .3 to .5 were assumed for cost modeling purposes. While an increase in the mobility parameter reduced the number of students that could be selected for a fixed cost, the impact on optimum cluster sizes and relative efficiencies for two versus three-stage designs were minimal. The Results reported in section IV of this paper indicate that 3.2 percent of the NLS seniors had changed counties by the second followup.

The topics presented in this report complete all of RTI's proposed methodological studies relating to the new cohort design. These studies have led to the following conclusions and recommendations:

1. Retain the deeply stratified two-stage design used for the 1972 cohort with the possible exception of type of control (public Catholic, and private non-Catholic) stratification. Selecting all but the smallest schools with probabilities proportional to estimated senior enrollment would reduce the effect of unequal weighting.

2. The 1972 design configuration with 1,200 schools and 18 sampled students per school is near optimal in terms of the total variable costs projected for a base-year and five followup surveys.

3. A matched sample retaining 50 percent of the 1972 schools for the new cohort could reduce the variance of new versus old cohort contrasts by up to 13 percent for national means. Procedures are available for using the most current school frame and the most up-to-date auxiliary data to draw the new sample and still achieve the desired level of overlap.

4. NCES should strongly consider dropping the use of substitute schools and students. Sample sizes can be preset to approximate desired levels of base-year response. Also, nonresponse bias can be partially compensated for by retrospective base-year responses obtained during the first followup and by weighting class imputations. The use of substitutes, however chosen, unnecessarily obscures final selection probabilities and complicates the assessment of response rates.

II. EFFECTS OF STRATIFICATION AND CLUSTERING

The effects of clustering the sample of students by school and the effect of stratifying schools have previously been approximated by Moore and Shah [1] using the base-year variance component estimates. These results are presented in table 1 for easy reference. In this section, these effects will be studied using change-variable variance components. Most of the formulas are developed in the paper by Moore and Shah [1]. However, a few will be repeated here for ease in following the analysis.

The estimated design effect (Deff) of a stratified two stage cluster sample can be expressed as

$$\text{Deff} = (n_2 \sigma_1^2 + \sigma_2^2)/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)$$

where

σ_0^2 is the variance component estimating the variation among final strata,

σ_1^2 is the variance component estimating the variation among schools within final strata,

σ_2^2 is the variance component estimating the variation among students within schools, and

n_2 is the average number of sample students per school.

If we define

$$\delta_{c/w} = (\sigma_0^2 + \sigma_1^2)/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)$$

and

$$\delta_{rs/w} = \sigma_0^2/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2),$$

then the design effect can be expressed by

$$Deff = 1 + (n_2 - 1)\ \delta_{c/w} - n_2\ \delta_{rs/w} \quad .$$

The term of these variance component ratios

$$C_{rw} = 1 + (n_2 - 1)\ \delta_{c/w}$$

represents the effect of clustering the sample of students by school attended, where $\delta_{c/w}$ is the intra-school cluster correlation for an unstratified selection of schools and students given the unequal weighting of the NLS sample design. The term, $n_2\delta_{rs/w}$, represents the reduction in the variances of survey estimates obtained from school stratification where $\delta_{rs/w}$ can be viewed as the intra-stratum cluster correlation appropriate for a design where clusters of students corresponding to entire final strata were selected and tested.

The effect of stratification, S_{cw}, in a multiplicative model can be expressed as

$$S_{cw} = (C_{rw} - n_2\delta_{rs/w})/C_{rw} \quad .$$

Then, the design effect, Deff, can be written as

$$Deff = C_{rw}S_{cw} \quad .$$

Thus, the design effect has been partitioned into C_{rw}, the effect of clustering,

and S_{cw}, the effect of stratification. The average effects of clustering and stratification, as well as the average values of $\delta_{c/w}$, $\delta_{rs/w}$ and the design effect, Deff, calculated for the NLS sample design are presented in table 1 using base-year variance component estimates and in table 2 using change-variable components. The S_{cw} values shown were derived from the average C_{rw} and Deff values. The increase in variance for an unstratified cluster sample over a simple random sample was 83.5% for base-year national means compared with 44.4% for change-variable national means. For subgroups the corresponding increase due to clustering varied from 18.8% to 65.5% for the base-year and from 17.3% to 35.9% for the change-variables with an average increase of 43.0% for the base-year compared with an average increase of 25.5% for the change-variables. For the all statistics group the increase was 39.1% for base-year and 23.6% for change-variables.

Stratifying the clusters using the NLS school stratification scheme reduced the variances for both the base-year and change-variable estimates. The percentage reduction is given by $100[1 - S_{cw}]$. This effect of stratification was greater for the base-year than for the change-variables for all subgroups except the domain whose fathers had more than a high school education. The reduction for national means was 20.3% for base-year and 12.9% for change-variables. For the subgroup means the reduction varied between 5.8% and 19.8% for base-year estimates and between 7.0% and 12.7% for change-variable estimates. The average reduction for all statistics was 14.3% for the base-year and 9.8% for the change-variables.

Table 1
Average effects of clustering and stratification for the NLS design using base-year variance component estimates*

Statistic	$\delta_{c/w}$	$C_{rw}^{\dagger}$	$\delta_{rs/w}$	$S_{cw}^{\dagger}$	$Deff^{\dagger}$
National means	.052	1.835	.022	.797	1.463
Subgroup means					
Whites	.041	1.655	.019	.802	1.327
Blacks	.029	1.458	.014	.836	1.219
Other races	.019	1.311	.008	.902	1.182
Females	.023	1.364	.009	.889	1.213
Males	.022	1.356	.010	.865	1.173
Father high school graduate	.019	1.302	.009	.888	1.156
Father less than high school	.017	1.274	.009	.877	1.117
Father more than high school	.012	1.188	.004	.942	1.119
All domain means	.027	1.430	.012	.836	1.233
Differences of domain and national means	.019	1.296	.009	.882	1.142
All statistics	.024	1.391	.011	.857	1.204

*All numbers in the table have been rounded to three digits after calculations. Substituting $\delta_{c/w}$ and $\delta_{rs/w}$ into formulas in text may result in small differences for C_{rw}, S_{cw}, and Deff.

†assumes $n_2 = 17$.

Table 2
Average effects of clustering and stratification for
the NLS design using change-variable variance component estimates*

Statistic	$\delta_{c/w}$	$C_{rw}^{\dagger}$	$\delta_{rs/w}$	$S_{cw}^{\dagger}$	$Deff^{\dagger}$
National means	.028	1.444	.011	.871	1.255
Subgroup means					
Whites	.022	1.359	.011	.873	1.180
Blacks	.017	1.273	.009	.872	1.111
Ocher races	.011	1.180	.005	.930	1.100
Females	.014	1.221	.005	.925	1.130
Males	.014	1.231	.007	.910	1.117
Father high school graduate	.011	1.173	.006	.909	1.073
Father less than high school	.011	1.174	.008	.888	1.044
Father more than high school	.011	1.177	.005	.926	1.094
All domain means	.016	1.255	.007	.899	1.130
Differences of domain and national means	.012	1.192	.006	.909	1.083
All statistics	.015	1.236	.007	.902	1.115

*All numbers in the table have been rounded to three digits after calculations.

†assumes $n_2 = 17$.

In general, the increase in variance due to cluster sampling is partially offset by the reduction due to stratification. The design effect (the column headed Deff in tables 1 and 2) measures the total effect of clustering and stratification. The total increase in variance as measured by the design effect was greater for the base-year estimates than for the change-variable estimates. The total increase for the national means was 46.3% for the base-year compared with 25.5% for change-variables. For subgroups, the total increase varied from 11.7% to 32.7% for base-year compared with an increase of 4.4% to 18.0% for change-variables. The average increase for the "all statistics" group was 20.4% for base-year as compared with 11.5% for change-variable estimates.

Design effects and variance estimates are also affected by the unequal weighting of the individual elements of the sample. The estimated design effects include the effect of unequal weighting, as do the estimated effects of stratification and clustering. Moore and Shah [1] conducted a limited and approximate analysis concerning the effects of unequal weighting. Their analysis was based on the sum of the estimated population in the low SES (socio-economic) stratum to the high SES stratum, the ratio of the sampling rates for the two strata, and the ratio of the proportions of the two strata belonging to a specified subgroup. Since the calculations are a function of the sample weights only, the results will be the same for both the base-year and change-variable variance estimates. The details of this study can be found in [1]. Moore and Shah concluded that sampling the low SES stratum at approximately twice the rate as for the high SES stratum resulted in an increase of about 13% in total variance compared with the variance estimate if the same sampling rate had been used for both strata. They also approximated

the increase in variance due to all causes of unequal weighting, such as nonresponse adjustments, unequal stratum sizes and imprecise school size measures. They estimated that unequal weighting increased the variance estimates by about 16% over the total sample with an increase of from 14% to 21% for the different subgroups.

III. CORRELATIONS FROM OVERLAPPING SCHOOLS IN THE TWO NLS SURVEYS

For measuring change in the means of two surveys taken at two time periods, the variance can often be reduced by using the same sampling units for both surveys. As compared to results from two independent samples, the variance of the differences between means is reduced to the degree that the means from overlapping sampling units (schools) are positively correlated. Generally some stability over time exists for most characteristics. Correlations between the same units for different time periods are almost always positive so that selecting overlapping units will result in a decrease in variance when comparing the two means. This section will determine the effect of this correlation on the difference in means between the 1972 NLS cohort and the new NLS cohort. For a more complete analysis of overlapping samples, see Kish, pp. 457-477 [2].

Let $\bar{x}$ be the estimated mean for the 1972 cohort, and $\bar{y}$ be the estimated mean for the new cohort. Then the difference between the two means is given by

$$d = \bar{y} - \bar{x} .$$

The variance of this difference can be expressed as

$$\text{var}(d) = \text{var}(\bar{y}) + \text{var}(\bar{x}) - 2\text{cov}(\bar{y},\bar{x}) .$$

If σ_1^2 is the estimated between school variance component, and σ_2^2 is the estimated between student within school variance component, then the variance of $\bar{x}$ and $\bar{y}$ is given as

$$\text{var}(\bar{x}) = \sigma_1^2(x)/n_1 + \sigma_2^2(x)/n_1 n_2$$

and $$\text{var}(\bar{y}) = \sigma_1^2(y)/m_1 + \sigma_2^2(y)/m_1 m_2$$

where n_1 is the number of schools selected for the 1972 cohort,

n_2 is the number of students selected per school for the 1972 cohort,

m_1 is the number of schools selected for the new NLS cohort, and

m_2 is the number of students selected per school for the new NLS cohort.

Note that only the overlapping schools will contribute to the $\text{cov}(\bar{x},\bar{y})$ since the student selection within schools will be independent from one survey to the next. The correlations in the covariance term from the overlapping schools refer only to the similarities of different students at different time periods within the same sample schools. The covariance term can be represented by

$$\text{cov}(\bar{x},\bar{y}) = \phi_1\sigma_1(xy)/n_1m_1 = [\phi_1 r_1(xy)\sigma_1(x)\sigma_1(y)]/n_1m_1$$

where ϕ_1 is the number of overlapping schools and $r_1(xy)$ is the correlation of student responses from the same school at the two time periods. The term $r_1(xy)$ is between -1 and 1 and its absolute value is given by

$$r_1(xy) = \sigma_1(xy)/\sigma_1(x)\sigma_1(y) .$$

Let $P_x = \phi_1/n_1$ be the proportion of 1972 sample schools that are also selected for the new sample of schools, and

$P_y = \phi_1/m_1$ be the proportion of sample schools for the new NLS cohort that were also 1972 sample schools.

The variance of the difference between the two means can then be expressed as

$$\text{var}(d) = [\sigma_1^2(y)/m_1 + \sigma_2^2(y)/m_1m_2] + [\sigma_1^2(x)/n_1 + \sigma_2^2(x)/n_1n_2]$$

$$- [2r_1(xy)P_xP_y\sigma_1(x)\sigma_1(y)/\phi_1] .$$

The overlapping of schools can be complete (the same schools in both samples), partial (P_x and P_y both less than one), or the two samples of schools can be completely independent. The above general formula will be used in simplified form to examine the variance of between cohort contrast for three different cases. Case 1 has no overlapping schools. In this case $\phi_1 = P_x = P_y = 0$. Case 2 has complete overlap of schools. In this case, $\phi_1 = n_1 = m_1$ and $P_x = P_y = 1$. Case 3 has incomplete overlap with the two samples containing the same number of sample schools and students. In this case, $n_1 = m_1 > \phi_1$, $n_2 = m_2$, and $P_x = P_y = P < 1$.

In order to estimate the reduction in variance that can be expected when estimating the difference between the means of the two NLS surveys, the following simplifying assumptions have been made:

$$\sigma_1^2(x) = \sigma_1^2(y) ,$$

$$\sigma_2^2(x) = \sigma_2^2(y) ,$$

$$n_1 = m_1 ,$$

$$n_2 = m_2 ,$$

$$P = P_x = P_y ,$$

and $$r = r_1(xy) .$$

The above assumptions permit simplification of the formula for calculating the variance of the difference between the two means. Table 3 summarizes the formulas for calculating the variance of the difference subject to the above assumptions.

Table 3. Formulas for calculating the variance of the differences between the two cohort means

	var(d)
Case 1: No overlap $r = 0$	$2\sigma_1^2/n_1 + 2\sigma_2^2/n_1n_2$
Case 2: Complete overlap $n_1 = m_1 = \phi_1$	$2\sigma_1^2(1-r)/n_1 + 2\sigma_2^2/n_1n_2$
Case 3: Partial overlap $n_1 = m_1 > \phi_1$	$2\sigma_1^2(1-rP)/n_1 + 2\sigma_2^2/n_1n_2$.

Note that cases 1 and 2 above are simplified versions of 3. If $P = 1$, case 3 results in case 2. When there is no overlap, $r = 0$, and case 3 becomes case 1.

Measuring to what extent overlapping schools will reduce the variance estimate for the new NLS cohort requires an estimate of σ_1^2, σ_2^2, and r. There are already estimates available for σ_1^2 and σ_2^2 from both the NLS base-year and change-variable data. Define the following variables

$$\delta_2 = (1 - \delta_{c/w})/(1 - \delta_{rs/w}) = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$$

and

$$\delta_1 = 1 - \delta_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2) ,$$

where $\delta_{c/w}$ and $\delta_{rs/w}$ are defined in section II. The percent reduction when comparing the estimated variance of a partial or complete overlap

to the case when there is no overlap is given by

$$100 \left\{ 1 - \frac{\text{variance estimate if partial or complete overlap}}{\text{variance estimate if no overlap}} \right\} .$$

The percent reduction in variance for both cases 2 and 3 over case 1 is given by

$$100 \left\{ 1 - \left(\frac{\sigma_1^2(1-rP) + \sigma_2^2/n_2}{\sigma_1^2 + \sigma_2^2/n_2} \right) \right\} .$$

In terms of NLS variance component ratios $\delta_{c/w}$ and $\delta_{rs/w}$, the percent reduction becomes

$$100 \left\{ 1 - \left(\frac{\delta_1(1-rP) + \delta_2/n_2}{\delta_1 + \delta_2/n_2} \right) \right\} ,$$

which reduces to $100\ [\delta_1 rP/(\delta_1 + \delta_2/n_2)]$.

When P = 1, case 2 is compared with case 1. The values of $\delta_{c/w}$ and $\delta_{rs/w}$ are given in table 1 using the base-year variance component estimates and in table 2 using the change-variable variance component estimates. The percent reduction that can be expected for three values of r (1, .75, and .50) and two values of P (.50 and .25) are presented in table 4 for base-year components. Average reductions have been computed for the following: national means, eight subgroup means, averages over all domain means, an average over all differences between domain and national means, and finally, an average of all statistics. The corresponding results for change-variable components are shown in table 5.

Overlapping schools decreased the variance more for base-year contrasts than for change-variable differences. As P, the proportion of overlapping schools, and r, the correlation coefficient decreased, the percent reduction in contrast

to the case when there is no overlap is given by

$$100 \left\{ 1 - \frac{\text{variance estimate if partial or complete overlap}}{\text{variance estimate if no overlap}} \right\} .$$

The percent reduction in variance for both cases 2 and 3 over case 1 is given by

$$100 \left\{ 1 - \left(\frac{\sigma_1^2(1-rP) + \sigma_2^2/n_2}{\sigma_1^2 + \sigma_2^2/n_2} \right) \right\} .$$

In terms of NLS variance component ratios $\delta_{c/w}$ and $\delta_{rs/w}$, the percent reduction becomes

$$100 \left\{ 1 - \left(\frac{\delta_1(1-rP) + \delta_2/n_2}{\delta_1 + \delta_2/n_2} \right) \right\} ,$$

which reduces to $100\ [\delta_1 rP/(\delta_1 + \delta_2/n_2)]$.

When P = 1, case 2 is compared with case 1. The values of $\delta_{c/w}$ and $\delta_{rs/w}$ are given in table 1 using the base-year variance component estimates and in table 2 using the change-variable variance component estimates. The percent reduction that can be expected for three values of r (1, .75, and .50) and two values of P (.50 and .25) are presented in table 4 for base-year components. Average reductions have been computed for the following: national means, eight subgroup means, averages over all domain means, an average over all differences between domain and national means, and finally, an average of all statistics. The corresponding results for change-variable components are shown in table 5.

Overlapping schools decreased the variance more for base-year contrasts than for change-variable differences. As P, the proportion of overlapping schools, and r, the correlation coefficient decreased, the percent reduction in contrast

Table 4
The percent reduction in the estimated variance when comparing means of two NLS cohorts due to overlapping schools using base-year variance component estimates*

Statistics	δ_1	δ_2	Percent reduction in variance**					
			r=1		r=.75		r=.50	
			P=.50	P=.25	P=.50	P=.25	P=.50	P=.25
National means	.031	.969	17.6	8.8	13.2	6.6	8.8	4.4
Subgroup means								
Whites	.022	.978	13.8	6.9	10.4	5.2	6.9	3.5
Blacks	.015	.985	10.3	5.1	7.7	3.9	5.1	2.6
Other races	.012	.988	8.6	4.3	6.4	3.2	4.3	2.1
Females	.014	.986	9.7	4.9	7.3	3.6	4.9	2.4
Males	.012	.988	8.6	4.3	6.4	3.2	4.3	2.1
Father H.S. grad	.010	.990	7.3	3.7	5.5	2.7	3.7	1.8
Father less than H.S.	.008	.992	6.0	3.0	4.5	2.3	3.0	1.5
Father more than H.S.	.003	.992	6.0	3.0	4.5	2.3	3.0	1.5
All domain means	.016	.984	10.8	5.4	8.1	4.1	5.4	2.7
Differences of domain and national means	.010	.990	7.3	3.7	5.5	2.7	3.7	1.8
All statistics	.013	.987	9.1	4.6	6.9	3.4	4.6	2.3

*Calculations were performed before rounding to three digits.
**Assume $n_2 = 17$.

The percent reduction in variance for P=1 can be obtained by doubling the corresponding values for P=.50 for each value of r.

Table 5
The percent reduction in the estimated variance when comparing means of two NLS cohorts due to overlapping schools using change-variable variance component estimates*

Statistics	δ_1	δ_2	Percent reduction in variance**					
			r=1		r=.75		r=.50	
			P=.50	P=.25	P=.50	P=.25	P=.50	P=.25
National means	.017	.983	11.4	5.7	8.5	4.3	5.7	2.8
Subgroup means								
Whites	.012	.988	8.6	4.3	6.4	3.2	4.3	2.1
Blacks	.008	.992	6.0	3.0	4.5	2.3	3.0	1.5
Other races	.007	.993	5.4	2.7	4.0	2.0	2.7	1.3
Females	.009	.991	6.7	3.3	5.0	2.5	3.3	1.7
Males	.008	.992	6.0	3.0	4.5	2.3	3.0	1.5
Father H.S. grad.	.005	.995	3.9	2.0	3.0	1.5	2.0	1.0
Father less than H.S.	.003	.997	2.4	1.2	1.8	0.9	1.2	0.6
Father more than H.S.	.006	.994	4.7	2.3	3.5	1.7	2.3	1.2
All domain means	.009	.991	6.7	3.3	5.0	2.5	3.3	1.7
Differences of domain and national means	.006	.994	4.7	2.3	3.5	1.7	2.3	1.2
All statistics	.008	.992	6.0	3.0	4.5	2.3	3.0	1.5

*Calculations were performed before rounding to three digits
**n_2 has been assumed equal to 17.

The percent reduction in variance for P=1 can be obtained by doubling the corresponding values for P=.50 for each value of r.

variance also decreased, as expected. However, for a 50 percent overlap and a value of r as low as .50, modest gains can be expected in the precision of the difference estimator. For the national means, a reduction of 8.8 percent for the base-year and 5.7 percent for the change-variables can be expected assuming r = .50 and P = .50. For the subgroup means, the reduction varies from 3.0 percent to 6.9 percent for the base-year and from 1.2 percent to 4.3 percent for the change-variables. For the all statistics category, this reduction is 4.6 percent for base-year and 3.0 percent for the change-variables.

Selecting some overlapping schools for the new NLS cohort will result in a decrease in variance for the comparison between the new cohort and the 1972 cohort. However, selecting overlapping schools for the new NLS cohort is not without its disadvantages.

The intital probabilities of selection of the 1972 sample schools were based on size measures obtained initially in 1972, which may diffe considerably from new measures in 1978 or when the new cohort is selected. Variations in the ratios of actual size to size measures cause variations in subsample sizes and tend to increase the variance of estimates for the new NLS cohort. Some variation in size can be tolerated but if this variation becomes too large, the size measures may need to be changed for the new survey. If measures of size are available for the schools at the time of selection of the new NLS cohort and if these differ considerably from those of the initial selection, then it is preferable to use the new size measures.

A method for changing from a set of initial probabilities p_j to a set of new probabilities P_j with a minimum number of actual changes of

sampling units was first presented by Keyfitz [3].

Assume there are D + I units in a stratum. Of these, D receive a decrease in probability, and I receive an increase, which may be zero. Let the subscripts distinguish the two sets of sampling units so that (d = 1, 2, ..., D) and (i = D + 1, D + 2, ..., D + I). Then $P_d < p_d$ and $P_i \geq p_i$.

The initial probabilities used for the selection of sampling units from a stratum, and the new probabilities both sum to unity in the stratum:

$$\sum_{d}^{D} p_d + \sum_{i}^{I} p_i = \sum_{d}^{D} P_d + \sum_{i}^{I} P_i \ .$$

The sum of the probability increases must equal the sum of the probability decreases:

$$\sum_{i}^{I} (P_i - p_i) = \sum_{d}^{D} (p_d - P_d) \ .$$

The rules for changing probabilities are:

(1) If the initially selected sampling unit shows an increase (or no change in probability), retain it in the sample, as if selected with the new probability P_i.

(2) If the initially selected unit shows a decrease, retain it in the sample with the probability P_d/p_d. Thus the compound probability of original selection plus retention is $p_d(P_d/p_d) = P_d$.

(3) If a decreasing unit is dropped from the sample, select a replacement among the increasing units, with probabilities proportional to their increases. The probability of selecting the ith unit is $(P_i-p_i)/\sum_{i}^{L} (P_i-p_i)$.

The total selection probability of an increasing unit consists of p_i initially, but is increased if any of the D decreasing units were selected and then dropped. The probability P_i is given by

$$P_i = p_i + \sum_{d}^{D} (p_d - P_d) \frac{P_i - p_i}{\sum_{i}^{I} (P_i - p_i)} .$$

The same changes or differences in population distributions that lead to changing selection probabilities can also motivate changing the strata. The best strata for the new population may be so different from the initial strata, that shifting the units is justified. In addition, new schools will need to be added and old schools deleted.

A simple procedure for the above can be the following:

(1) All schools in a new stratum that did not belong to it initially are treated as newly created with $p_j = 0$. Some new strata may consist entirely of such new units.

(2) All schools that belonged originally but are no longer present in the stratum are eliminated with $P_j = 0$. Rules are needed to identify uniquely each of the initial strata with only one of the new strata. The rules must avoid selection bias, and should maximize retention of initial sample schools.

There exist modifications of the above procedure. See Keyfitz [3] and Kish [2], for example. The above description is very brief and is given merely as an illustration of the type of procedure recommended for selecting the new cohort sample of overlapping schools. Additional refinements relating to without replacement selection would also be required for the new NLS cohort schools.

In conclusion, the gains in precision due to some overlap of sample schools is sufficient for RTI to recommend this procedure. It would appear that no more than 50 percent of the schools should overlap. A procedure similar to the method presented by Keyfitz should be used to update the initial school selection probabilities and to form new strata if needed to minimize the variance for estimates for the new NLS cohort. The use of a procedure to update the initial school selection probabilities complicates the school selection process and increases the school sampling costs somewhat. However, this increase should be a small portion of the total budget and may be compensated by a decrease in data collection due to the overlapping schools.

IV. MOBILITY OF THE NLS 1972 COHORT

The mobility of sample students has direct cost implications for NLS followup surveys. Students who do not respond to the mail questionnaires must be interviewed personally. Interviewer salaries and travel expenses represent a sizeable portion of the data collection costs. In the two-stage three-stage design comparison reported by Folsom and Weber [5], the mobility of the sample students influenced the interviewer-travel model and in turn the cost components for the two- and three-stage designs. The relative efficiencies of a three-versus a two-stage design were not affected appreciably by student outmigration from their base-year county in the range of 30% to 50%. The total number of students that can be selected for a fixed cost does depend on the number expected to move as well as the number expected to respond to the mail questionnaire. The mail response rate has been well documented. In this section the number of sample students who are considered movers from the base-year (BY) to the first followup (FFU) and from BY to second followup (FSU) are presented.

A student's base-year county is defined by using the zip code address of the student's school. The set of counties overlapped by the school's zip code constitute the base-year county or counties for all sample students from that school. A student is considered to be a mover if the county or counties overlapped by his first or second followup zip code address is different than his base-year county.

A state/county 1976 geographic data file tape was obtained from List Processing Company, Incorporated. Zip code addresses from each sample student's BY, FFU, and SFU records were matched with zip codes in each county

on the state/county data tapes to determine BY, FFU, and SFU counties. Several problems were encountered while trying to match base-year, first followup, and second followup counties. There were addresses on the student file having no zip codes. In addition, by the first or second followup there were no addresses available for some of the sample students. For a few records an address with a zip code was available, but the zip code did not match the zip codes in any of the counties listed on the state/county data tape. Two possible explanations for nonmatching zip codes are:

(a) Out-of-date zip codes, and
(b) Coding errors for either the NLS file or the state/county data tape.

Table 6 lists the counts of records in each of the possible categories.

Table 6. Number of student records by matching category for BY, FFU, and SFU

	BY	FFU	SFU
Address, but no zip code	0	3,232	376
No address available	0	135	2,579
Address and zip code available but did not match state/county data tape	319	422	539
Zip code available and matched to county	23,132	19,662	19,957
Total records	23,451	23,451	23,451

Note that for the FFU there were 3,232 addresses with no associated zip code while at the SFU there were only 376 such addresses. This reduction in missing zips was largely offset by an increase in the number of students for whom no address was available. To determine if a student was a "mover,"

a zip code to county match was required for both the associated BY and FFU or SFU addresses. There were 19,259 student records where zip code to county matches were found for both BY and FFU addresses. Of these 19,259 students, 5,389 were "movers"; that is, 28.0 percent of the students comparisons could be made between the BY and FFU surveys. There were 19,497 student records with a county match for both their BY and SFU zip codes. Among these 19,497 address pairs 6,473 county shifts were observed for a mover percentage of 33.2 percent between the BY and SFU. Of the 17,751 students where both their FFU and SFU zips had county matches, 3,500 (19.7 percent) changed counties.

The sum of "movers" from BY to FFU and from FFU to SFU exceeds the number of "movers" from BY to SFU. This is not altogether surprising since students who go away to college often drop out after a year or two and return home. Other graduates move away to find work and then return. In spite of this return migration, the net outmigration would be expected to increase as each followup progresses. The observed rate of increase between the first and second followups was around five percentage points.

The mobility rates computed here are consistent with the range of values previously assumed when modeling costs for alternative two- and three-stage designs [4] and [5]. Outmigration between the BY and SFU could be as high as 41 percent if all of the 2,579 students with unknown addresses at the second followup had moved away. This proportion would still fall within the 30 percent to 50 percent mover range hypothesized.

In conclusion, the observed proportions of sample students moving from BY to FFU and from BY to SFU are consistent with the range of values hypothesized in previous reports. These results support the earlier optimizations which suggest that the 1972 cohort design with 1,200 schools and 18 sample students per school is near optimal.

REFERENCES

[1] R. P. Moore and B. V. Shah, "Effects of Stratification, Clustering, and Unequal Weighting on the Variances of NLS Statistics," Report on RTI Project 22U-884-3, Research Triangle Institute, Research Triangle Park, N.C., February 1975.

[2] Leslie Kish, Survey Sampling. John Wiley and Sons, New York, 1965.

[3] N. Keyfitz, "Sampling with Probability Proportional to Size; Adjustment for Changes in Probabilities," JASA, 46, 105-109.

[4] R. P. Moore and B. V. Shah, "Efficiency Study of the Change Variables from the First Follow-up of the NLS Sample Design," Report on RTI Project 22U-884-21, Research Triangle Institute, Research Triangle Park, N.C., February 1976.

[5] R. E. Folsom and J. H. Weber, "Relative Efficiencies of a Three-stage Versus a Two-stage Sample Design for a New NLS Cohort Study," Report on RTI Project 22U-884-38, Research Triangle Institute, Research Triangle Park, N.C., March 1976.